

# Assignment 11

*Christophe Hunt*

*April 23, 2017*

## Contents

|          |   |          |
|----------|---|----------|
| <b>1</b> | <b>Perform a linear regression analysis fitting the Max Heart Rate to Age using the lm function in R.</b>   | <b>2</b> |
| 1.1      | What is the resulting equation? . . . . .   | 2        |
| 1.2      | Is the effect of Age on Max HR significant? . . . . .   | 2        |
| 1.3      | What is the significance level? . . . . .   | 3        |
| 1.4      | Please also plot the fitted relationship between Max HR and Age. . . . .  | 3        |
| <b>2</b> | <b>What is the final linear regression fit equation Using the Auto data set from Assignment 5?</b>  | <b>4</b> |
| 2.1      | Which of the 4 independent variables have a significant impact on mpg? . . . . .  | 5        |
| 2.2      | What are their corresponding significance levels? . . . . .   | 5        |
| 2.3      | What are the standard errors on each of the coefficients? . . . . .   | 5        |
| <b>3</b> | <b>Please perform this experiment in two ways.</b>  | <b>6</b> |
| 3.1      | First take any random 40 data points from the entire auto data sample and perform the linear regression fit and measure the 95% confidence intervals. . . . . | 6        |
| 3.2      | Then, take the entire data set (all 392 points) and perform linear regression and measure the 95% confidence intervals. . . . .                               | 8        |

Using R's lm function, perform regression analysis and measure the significance of the independent variables for the following two data sets.

In the first case, you are evaluating the statement that we hear that Maximum Heart Rate of a person is related to their age by the following equation:

$$\text{MaxHR} = 220 - \text{Age}$$

You have been given the following sample:

Age 18 23 25 35 65 54 34 56 72 19 23 42 18 39 37

MaxHR 202 186 187 180 156 169 174 172 153 199 193 174 198 183 178

# 1 Perform a linear regression analysis fitting the Max Heart Rate to Age using the lm function in R.

## 1.1 What is the resulting equation?

```
library(stargazer)
Age <- c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37)
MaxHR <- c(202, 186, 187, 180, 156, 169, 174, 172, 153, 199, 193, 174, 198, 183, 178)
fit <- lm(MaxHR~Age)
stargazer(fit, header = FALSE)
```

| Table 1:                |                             |
|-------------------------|-----------------------------|
|                         | <i>Dependent variable:</i>  |
|                         | MaxHR                       |
| Age                     | −0.798***<br>(0.070)        |
| Constant                | 210.048***<br>(2.867)       |
| Observations            | 15                          |
| R <sup>2</sup>          | 0.909                       |
| Adjusted R <sup>2</sup> | 0.902                       |
| Residual Std. Error     | 4.578 (df = 13)             |
| F Statistic             | 130.009*** (df = 1; 13)     |
| Note:                   | *p<0.1; **p<0.05; ***p<0.01 |

## 1.2 Is the effect of Age on Max HR significant?

Yes, the effect of Age on MaxH is significant. We see the p<0.01 for Age which indicates that it is indeed significant. We would say anything with a p<0.05 is typically significant, depending on the situation.

### 1.3 What is the significance level?

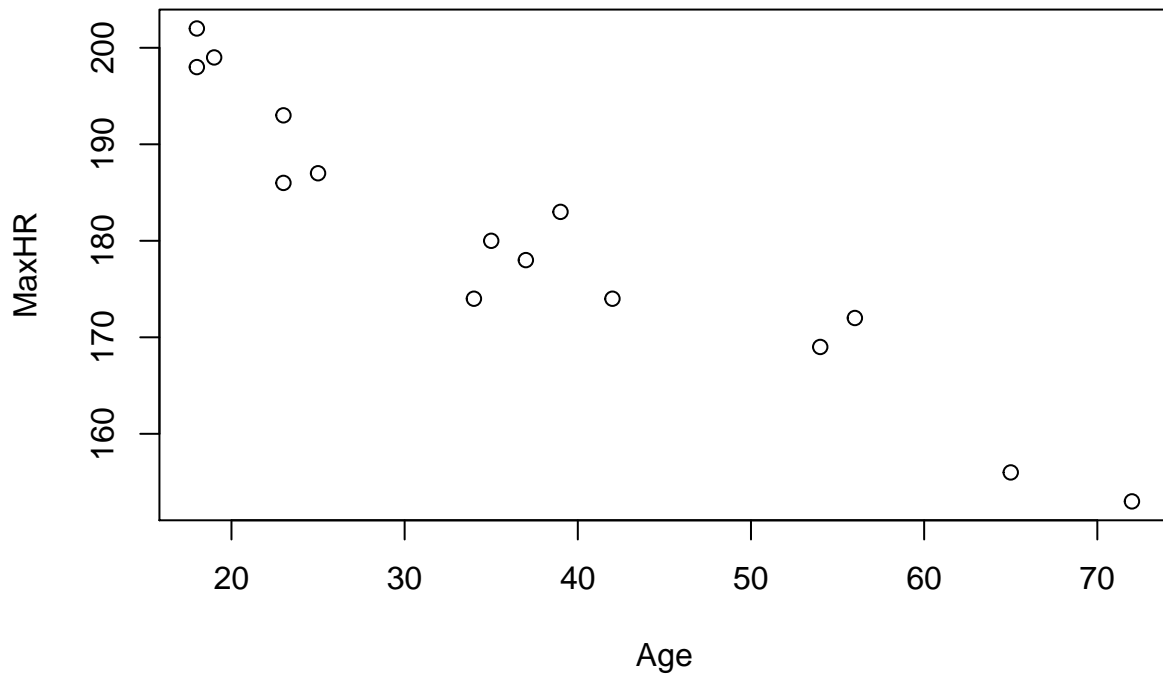
The significance level is  $p < 0.01$ , the exact value is as follows.

```
options(scipen = 999)
x <- anova(fit)
x$`Pr(>F)`[1]
```

```
## [1] 0.00000003847987
```

### 1.4 Please also plot the fitted relationship between Max HR and Age.

```
plot(MaxHR~Age)
```



We can see a very clear linear relationship between the two variables. This helps in confirming that a linear model should be appropriate for this data set.

Using the Auto data set from Assignment 5 (also attached here) perform a Linear Regression analysis using mpg as the dependent variable and the other 4 (displacement, horsepower, weight, acceleration) as independent variables.

## 2 What is the final linear regression fit equation Using the Auto data set from Assignment 5?

```
library(tidyverse)
df <- as_tibble(read.table(paste0("https://raw.githubusercontent.com",
                                "/ChristopheHunt/MSDA---Coursework",
                                "/master/Data%20605/Assignment%2011/",
                                "auto-mpg.data")))
colnames(df) <- c("displacement", "horsepower", "weight", "acceleration", "mpg")

fit <- lm(data = df, formula = (mpg ~ displacement + horsepower +
                                weight + acceleration))
stargazer(fit, header = FALSE)
```

| Table 2:                |  |
|-------------------------|--|
|                         | <i>Dependent variable:</i>                             |
|                         | mpg ~displacement + horsepower + weight + acceleration |
| displacement            | −0.006<br>(0.007)                                      |
| horsepower              | −0.044***<br>(0.017)                                   |
| weight                  | −0.005***<br>(0.001)                                   |
| acceleration            | −0.023<br>(0.126)                                      |
| Constant                | 45.251***<br>(2.456)                                   |
| Observations            | 392  |
| R <sup>2</sup>          | 0.707  |
| Adjusted R <sup>2</sup> | 0.704  |
| Residual Std. Error     | 4.247 (df = 387)                                       |
| F Statistic             | 233.434*** (df = 4; 387)                               |
| Note:                   | *p<0.1; **p<0.05; ***p<0.01                            |

## 2.1 Which of the 4 independent variables have a significant impact on mpg?

Horsepower and weight have a significant impact on mpg. It is unlikely that their values have an impact on mpg merely by chance.

## 2.2 What are their corresponding significance levels?

```
summary(fit)[4]
```

```
## $coefficients
##           Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept)  45.251139699  2.4560446927  18.4243959  7.072099e-55
## displacement -0.006000871  0.0067093055  -0.8944102  3.716584e-01
## horsepower   -0.043607731  0.0165734633  -2.6311779  8.848982e-03
## weight       -0.005280508  0.0008108541  -6.5122789  2.302545e-10
## acceleration -0.023147999  0.1256011622  -0.1842977  8.538765e-01
```

The variables corresponding significance levels are horsepower = 0.009 & weight = 0.0000000002

## 2.3 What are the standard errors on each of the coefficients?

```
summary(fit)[4]$coefficients
```

```
##           Estimate   Std. Error   t value   Pr(>|t|)
## (Intercept)  45.251139699  2.4560446927  18.4243959  7.072099e-55
## displacement -0.006000871  0.0067093055  -0.8944102  3.716584e-01
## horsepower   -0.043607731  0.0165734633  -2.6311779  8.848982e-03
## weight       -0.005280508  0.0008108541  -6.5122789  2.302545e-10
## acceleration -0.023147999  0.1256011622  -0.1842977  8.538765e-01
```

The variables corresponding standard errors for our coefficients are horsepower = 0.0166 & weight = 0.00081

### 3 Please perform this experiment in two ways.

#### 3.1 First take any random 40 data points from the entire auto data sample and perform the linear regression fit and measure the 95% confidence intervals.

```
set.seed(1234)
df_sample <- sample_n(df, 40)
fit.samp <- lm(data = df_sample, formula = (mpg ~ displacement +
                                           horsepower + weight + acceleration))
stargazer(fit.samp, header = FALSE)
```

Table 3:

|                         | Dependent variable:                                    |
|-------------------------|--|
|                         | mpg ~displacement + horsepower + weight + acceleration |
| displacement            | 0.010<br>(0.030)                                       |
| horsepower              | -0.177**<br>(0.078)                                    |
| weight                  | -0.002<br>(0.004)                                      |
| acceleration            | -0.413<br>(0.498)                                      |
| Constant                | 52.995***<br>(8.262)                                   |
| Observations            | 40   |
| R <sup>2</sup>          | 0.710  |
| Adjusted R <sup>2</sup> | 0.677  |
| Residual Std. Error     | 4.595 (df = 35)  |
| F Statistic             | 21.396*** (df = 4; 35)                                 |

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

```
names <- colnames(df)[1:4]
for (i in names){
  kable(print(confint(fit.samp, i, level = 0.95)))
}
```

```
##           2.5 %      97.5 %
## displacement -0.0509389 0.07008803
##           2.5 %      97.5 %
## horsepower -0.3361474 -0.01804589
##           2.5 %      97.5 %
## weight -0.01142215 0.006696022
##           2.5 %      97.5 %
## acceleration -1.423285 0.596921
```

You'll notice that the values that are not significant have confidence intervals that cross 0. If the

value could be 0, then its possible that the independent variable impacts our dependent variable merely by chance.

### 3.2 Then, take the entire data set (all 392 points) and perform linear regression and measure the 95% confidence intervals.

```
fit <- lm(data = df, formula = (mpg ~ displacement +
                               horsepower + weight + acceleration))
stargazer(fit, header = FALSE)
```

Table 4:

|                         | Dependent variable:                                    |
|-------------------------|--|
|                         | mpg ~displacement + horsepower + weight + acceleration |
| displacement            | −0.006<br>(0.007)                                      |
| horsepower              | −0.044***<br>(0.017)                                   |
| weight                  | −0.005***<br>(0.001)                                   |
| acceleration            | −0.023<br>(0.126)                                      |
| Constant                | 45.251***<br>(2.456)                                   |
| Observations            | 392  |
| R <sup>2</sup>          | 0.707  |
| Adjusted R <sup>2</sup> | 0.704  |
| Residual Std. Error     | 4.247 (df = 387)                                       |
| F Statistic             | 233.434*** (df = 4; 387)                               |
| Note:                   | *p<0.1; **p<0.05; ***p<0.01                            |

```
names <- colnames(df)[1:4]
for (i in names){
  kable(print(confint(fit, i, level = 0.95)))
}
```

```
##           2.5 %      97.5 %
## displacement -0.01919212 0.00719038
##           2.5 %      97.5 %
## horsepower -0.07619303 -0.01102243
##           2.5 %      97.5 %
## weight -0.006874738 -0.003686277
##           2.5 %      97.5 %
## acceleration -0.270094 0.2237981
```



Let's compare the results of the model using the larger data set and the smaller 40 case sample.

```
names <- colnames(df)[1:4]
for (i in names){
  kable(print(confint(fit, i, level = 0.95)))
  kable(print(confint(fit.samp, i, level = 0.95)))
}
```

```
##           2.5 %      97.5 %
## displacement -0.01919212 0.00719038
##           2.5 %      97.5 %
## displacement -0.0509389 0.07008803
##           2.5 %      97.5 %
## horsepower -0.07619303 -0.01102243
##           2.5 %      97.5 %
## horsepower -0.3361474 -0.01804589
##           2.5 %      97.5 %
## weight -0.006874738 -0.003686277
##           2.5 %      97.5 %
## weight -0.01142215 0.006696022
##           2.5 %      97.5 %
## acceleration -0.270094 0.2237981
##           2.5 %      97.5 %
## acceleration -1.423285 0.596921
```

It is interesting to see how the smaller data set impacts our results. It highlights how important each observation can be in obtaining significant results. It also highlights why we reject the null hypothesis rather than accept the new hypothesis because we can rarely be certain an independent variable is the cause of the change in our dependent variable. A small observation group can give us erroneous results. Hypothesis testing is a powerful tool and one should be aware of the possibility in making Type I & Type II errors.